ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Computer and System Sciences 70 (2005) 539–554

JOURNAL OF COMPUTER AND SYSTEM SCIENCES

www.elsevier.com/locate/jcss

Predictive complexity and information

Michael V. Vyugin[a], Vladimir V. V'yugin[b,*]

[a] *Department of Computer Science, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK*

[b] *Institute for Information Transmission Problems, Russian Academy of Sciences, Bol'shoi Karetnyi per. 19, Moscow GSP-4, 101447, Russia*

Received 8 January 2003; received in revised form 13 May 2003

Available online 1 December 2004

Abstract

The notions of predictive complexity and of corresponding amount of information are considered. Predictive complexity is a generalization of Kolmogorov complexity which bounds the ability of any algorithm to predict elements of a sequence of outcomes. We consider predictive complexity for a wide class of bounded loss functions which are generalizations of square-loss function. Relations between unconditional $\mathrm{KG}(x)$ and conditional $\mathrm{KG}(x|y)$ predictive complexities are studied. We define an algorithm which has some "expanding property". It transforms with positive probability sequences of given predictive complexity into sequences of essentially bigger predictive complexity. A concept of amount of predictive information $\mathrm{IG}(y : x)$ is studied. We show that this information is noncommutative in a very strong sense and present asymptotic relations between values $\mathrm{IG}(y : x)$, $\mathrm{IG}(x : y)$, $\mathrm{KG}(x)$ and $\mathrm{KG}(y)$.

Keywords: Machine learning; Algorithmic prediction; Predictive complexity; Predictive information; Loss functions; Kolmogorov complexity; Expanding property

1. Introduction

Machine learning (and statistics) considers the problem of predicting a future event x_n from past observations $x_1x_2\ldots x_{n-1}$, where $n = 1, 2\ldots$. A prediction algorithm makes its prediction in a

* Corresponding author.

E-mail addresses: misha@cs.rhul.ac.uk (M.V. Vyugin), vyugin@iitp.ru (V.V. V'yugin).

0022-0000/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcss.2004.10.005

form of a real number between 0 and 1. The quality of prediction is measured by a loss function $\lambda(\sigma, p)$, where σ is an outcome and $0 \leqslant p \leqslant 1$ is a prediction. Various loss functions are considered in literature on machine learning and prediction with expert advice [1,3,8,9,14]. The logarithmic loss function, where $\lambda(\sigma, p) = -\log p$ if $\sigma = 1$ and $\lambda(\sigma, p) = -\log(1-p)$ otherwise, is important in probability theory and generates the log-likelihood function. The square-loss function $\lambda(\sigma, p) = (\sigma - p)^2$ is important for applications.

The main goal of on-line prediction is to find a method of prediction which minimizes the total loss suffered on a sequence $x = x_1 x_2 \ldots x_n$ for $n = 1, 2 \ldots$. This idea of "minimal" possible total loss was formalized by Vovk [10] in a notion of predictive complexity. Intuitively, it is the loss of an optimal idealized prediction strategy ("idealized" means that this strategy is allowed to work infinitely long). This loss gives a natural lower bound to ability of any algorithm to predict elements of a sequence of outcomes. For the exact definition of predictive complexity see Section 2. Since predictive complexity is defined only up to an additive constant, it cannot have direct applications in practice. However it allows us to give a very general formal representation to our intuitions about prediction. It provides a tool to determine the limits of effectiveness of prediction algorithms. An example of an application of this tool is the statement and the solution of the snooping problem published in [11]: given a bit of information about a data set, how much better can elements of this data set be predicted?

The situation is similar to that with Kolmogorov complexity, which formalizes our intuitions about the shortest description of an object. Results of the theory of predictive complexity are expressed in the same asymptotic fashion as the results about Kolmogorov complexity. In fact, Kolmogorov complexity coincides with predictive complexity for a particular loss function.

In this paper, we present results for a class of bounded loss functions which are generalizations of the squared difference. We study relations between unconditional and conditional predictive complexities. The motivation for this study can be provided by the following practical problem. Suppose we have to predict a sequence of data $y = y_1, y_2, y_3, \ldots$ (say, prices of houses) given another sequence on-line $z = z_1, z_2, z_3, \ldots$ (say, a set of attributes of these houses) and there is third sequence $x = x_1, x_2, x_3, \ldots$ (which contains some other attribute). The question is: given sequences z and x on-line, how much better can sequence y be predicted than in the situation when only z is given? The sequence z can be considered as an oracle and all further results can be rewritten for the oracle case. Thus we will not mention z below. This problem justifies our interest in the comparison of minimal losses on y and on y given x, or in other words $\mathrm{KG}(y)$ and $\mathrm{KG}(y|x)$. The value $\mathrm{KG}(x)$ is also important. A difference between $\mathrm{KG}(y)$ and $\mathrm{KG}(y|x)$ can be considered as an amount of information in x about y. We denote it as $\mathrm{IG}(x : y)$ and call *amount of predictive information*.

At first, we prove a variant of triangle inequality for predictive complexity (Proposition 4). The main result of the paper, Theorem 2, implies that the upper bound in this inequality is tight for a wide range of cases. In order to prove it, we describe an algorithm Φ (Theorem 1) which has some "expanding property": this algorithm transforms with positive probability a sequence x of given predictive complexity k into a sequence $\Phi(x)$ with predictive complexity $\mathrm{KG}(\Phi(x)) \geqslant ck \log \frac{n}{k}$, where n is the length of x, c is a constant. This is impossible for Kolmogorov complexity $K(x)$, since for any computable mapping Φ, it satisfies the inequality $K(\Phi(x)) \leqslant K(x) + O(1)$. Theorem 3 is an interesting reformulation of the main result showing that pair of sequences can be predicted easier than its element.

The second group of results concerning predictive information is given in Section 4. We explore relations between four important values $\mathrm{IG}(y : x)$, $\mathrm{IG}(x : y)$, $\mathrm{KG}(x)$ and $\mathrm{KG}(y)$ in a limit form (Theorems 4–7). In particular, we prove that $\mathrm{IG}(y : x)$ is noncommutative in a very strong sense.

2. Predictive complexity

We consider only simplest case, where events are simple binary outcomes from $\{0, 1\}$. Let us denote $\Xi_n = \{0, 1\}^n$ and $\Xi = \bigcup_{n=0}^{\infty} \Xi_n$. We denote by $l(x)$ the length of a finite sequence $x \in \Xi$, $x^n = x_1 \dots x_n$ is its initial prefix of length n. By ε we denote the empty sequence from Ξ. We write $x \subseteq y$ if $x = y^n$ for some $n \leqslant l(y)$. We also will identify elements of Ξ and positive integer numbers in a natural fashion. It is important to consider the case when some a priori information is used in performing predictions. We consider a set of *signals* Δ. For simplicity we will consider the case $\Delta = \Xi$.

It is natural to suppose that all predictions are given according to a *prediction strategy* (or *prediction algorithm*) S. When performing prediction p_i the strategy S uses two input sequences, a sequence $x^{i-1} = x_1 x_2 \dots x_{i-1}$ of previous outcomes, and a sequence $y^i = y_1 y_2 \dots y_i$ of signals, i.e. $p_i = S(x^{i-1}, y^i)$. The total loss suffered by S on a sequence x^n is represented as

$$\mathrm{Loss}_S(x^n|y^n) = \sum_{i=1}^{n} \lambda(x_i, S(x^{i-1}, y^i)).$$

The value $y^n = 0^n$, $n = 0, 1, \dots$, corresponds to the case when a priori information does not used (here by 0^n we denote the sequence of n zeros).

The value $\mathrm{Loss}_S(x^n|y^n)$ can be interpreted as predictive complexity of x^n given y^n. This value, however, depends on S and it is unclear which S to choose. Levin [15], developing ideas of Kolmogorov and Solomonoff, suggested (for the logarithmic loss function) a very natural solution to the problem of existence of a smallest measure of predictive complexity. Vovk [10] extended these ideas in a more general setting for the class of mixable loss functions.

Let us fix $\eta > 0$ (the *learning rate*) and put $\beta = e^{-\eta} \in (0, 1)$. A loss function (game) $\lambda(\sigma, \gamma)$ is called η-mixable if for any sequence of predictions $\gamma_1, \gamma_2, \dots$ and for any sequence of nonnegative weights $p_1, p_2, \dots$ with sum $\leqslant 1$ a prediction $\hat{\gamma}$ exists such that

$$\lambda(\sigma, \hat{\gamma}) \leqslant \log_\beta \sum_i p_i \beta^{\lambda(\sigma, \gamma_i)} \tag{1}$$

for all σ. The log-loss function is η-mixable for any $0 < \eta \leqslant \ln 2$, and square loss function is also η-mixable for any $0 < \eta \leqslant 2$ (see [8]).

A function $\mathrm{KG}(x|y)$ is a *measure of predictive complexity* (with respect to a mixable loss function $\lambda(\sigma, \gamma)$) if the following two conditions hold:

- (i) $\mathrm{KG}(\varepsilon|\varepsilon) = 0$ and for every x and y of equal length and each extension γ of y there exists a prediction p depending on x and $y\gamma$ such that inequality

$$\mathrm{KG}(x\sigma|y\gamma) \geqslant \mathrm{KG}(x|y) + \lambda(\sigma, p) \tag{2}$$

  holds for each σ.
- (ii) KG is *semicomputable from above*, which means that there exists a computable nondecreasing by t sequence of simple functions KG^t such that for every x and y it holds $\mathrm{KG}(x|y) = \inf_t \mathrm{KG}^t(x|y)$.

By a simple function we mean a function which takes rational values or $+\infty$ and equals $+\infty$ for almost all $x \in \Xi$.

Requirement (i) means that the measure of predictive complexity must be valid: there exists a prediction strategy that achieves it. Notice that if $\geqslant$ is replaced by $=$ in (2), then the definition of a total loss function will be obtained. Requirement (ii) means that $\mathrm{KG}(x|y)$ must be "computable in the limit".

The main advantage of such definition is that a semicomputable from above sequence $\mathrm{KG}_i(x|y)$ of all measures of predictive complexity (with respect to a fixed mixable loss function $\lambda(\sigma,\gamma)$) can be constructed. More precise, there exists a computable from i, t, x, y sequence $\mathrm{KG}_i^t(x|y)$ of simple functions such that

- (iii) $\mathrm{KG}_i^{t+1}(x|y) \leqslant \mathrm{KG}_i^t(x|y)$ for all i, t, x;
- (iv) $\mathrm{KG}_i(x|y) = \inf_t \mathrm{KG}_i^t(x|y)$ for all i, x;
- (v) for each measure of predictive complexity $\mathrm{KG}(x|y)$ there exists an i such that $\mathrm{KG}(x|y) = \mathrm{KG}_i(x|y)$ for all x and y. In particular, for any computable prediction strategy S there exists an i such that $\mathrm{Loss}_S(x|y) = \mathrm{KG}_i(x|y)$ for all x and y.

This sequence can be constructed using standard methods of the theory of algorithms [6]. Exact construction and proofs for unconditional case are given in [12].

Let us consider some analogy with Kolomogorov complexity which is based on computable methods of decoding of finite binary sequences. By this method F of decoding we can reconstruct any finite sequence x using its binary program p and some parameter y: $x = F(p, y)$. Each method of decoding F defines some measure of conditional complexity $K_F(x|y) = \min\{l(p) : F(p, y) = x\}$ of finite sequences x. It is easy to verify that this function is semicomputable from above. Kolmogorov's idea was to "mix" all these measures of complexity in a "universal" measure. A computable sequence F_i of all methods of decoding can be constructed by the methods of the theory of algorithms [6]. A universal method of decoding can be defined $U(\langle i, p, y\rangle) = F_i(p, y)$, where i is a program computing F_i and $\langle i, p, y\rangle$ is a suitable code of the triple (i, p, y). Then for any method of decoding F semicomputable from above it holds $K_U(x|y) \leqslant K_F(x|y) + O(1)$ for each x and y, where a constant $O(1)$ depends on F. We fix some $K_U(x|y)$, denote it $K(x|y)$, and call the conditional Kolmogorov complexity of a finite sequence x given a finite sequence y. Unconditional complexity is defined $K(x) = K(x|\varepsilon)$. For technical reason we consider prefix-free methods of decoding: for any (p, y) and (p', y') from the domain of F if $p \neq p'$ then $p \not\sqsubseteq p'$ and $p' \not\sqsubseteq p$. Then $\sum 2^{-K(x)} \leqslant 1$. In prefix-free case Kolmogorov complexity can be also defined analytically (see [5]).

$$K_U(x|y) = \log_{1/2} \sum_{i=1}^{\infty} r_i 2^{-K_{F_i}(x|y)},$$

where $r_1, r_2, \ldots$ is a computable sequence of nonnegative weights with sum $\leqslant 1$. For example, we can take $r_i = 2^{-i}$, $i = 1, 2, \ldots$.

The mixture of all measures of predictive complexity $\mathrm{KG}_i(x|y)$ in the case of η-mixable loss function is defined

$$\mathrm{KG}(x|y) = \log_\beta \sum_{i=1}^{\infty} r_i \beta^{\mathrm{KG}_i(x|y)}, \tag{3}$$

where $r_i = 2^{-K(i)}$, $i = 1, 2, \ldots$. The following proposition asserts that the function $\mathrm{KG}(x|y)$ is a measure of predictive complexity minimal up to an additive constant.

Proposition 1. *Let a loss function $\lambda(\omega, p)$ be computable and η-mixable for some $\eta > 0$. Then there exists a measure of predictive complexity* $\mathrm{KG}(x|y)$ *with respect to $\lambda(\sigma, \gamma)$ such that for each measure of predictive complexity* $\mathrm{KG}_i(x|y)$ (*with respect to $\lambda(\sigma, \gamma)$*)

$$\mathrm{KG}(x|y) \leqslant \mathrm{KG}_i(x|y) + (\ln 2/\eta)K(i) \tag{4}$$

holds for all x, y, where $K(i)$ is the Kolmogorov prefix complexity of the program i enumerating KG_i *from above, besides, a constant c exists such that* $\mathrm{KG}(x|y) \leqslant \mathrm{Loss}_S(x|y) + (\ln 2/\eta)(K(S) + c)$ *holds for each computable prediction strategy S and for each x, y.* [1]

The proof of this proposition is based on Vovk's aggregating algorithm [7,8,10] (see also Section A).

Let some η-mixable loss function $\lambda(\sigma, p)$ be given. Put

$$b = \inf_p \sup_\sigma \lambda(\sigma, p). \tag{5}$$

We suppose that $b > 0$. We suppose also that the loss function is computable, and hence, it is continuous in p in the interval $[0, 1]$. We also suppose that the infinum in (5) is attained for some computable real number $\hat{p}$. For log-loss function $b = 1$, $p = \frac{1}{2}$ and $b = 1/4$, $p = \frac{1}{2}$ in the case of squared difference.

We impose a very natural condition

$$\lambda(0, 0) = \lambda(1, 1) = 0, \tag{6}$$

which holds in both cases of log- and square-loss functions. We also consider additional requirements on loss function by which square-loss function differs from log-loss function

$$\lambda(0, 1) = \lambda(1, 0) = a. \tag{7}$$

Now, when restrictions on a loss function were specified let us fix some $\mathrm{KG}(x|y)$ satisfying conditions of Proposition 1 and call its value the conditional *predictive complexity* of x given y. In the case when y is trivial, i.e. consists only from zeros we consider (unconditional) predictive complexity $\mathrm{KG}(x)$ of a sequence x.

Let us consider the set

$$A_{n,k} = \{y | l(y) = n, \mathrm{KG}(y) \leqslant k\}. \tag{8}$$

We denote by $\#A$ the cardinality of a finite set A. Tight upper and lower bounds of the number $\#A_{n,k}$ of all sequences of length n having predictive complexity $\leqslant k$ were obtained in [12]. Relations between Kolmogorov and predictive complexities also were obtained in that paper. Here, we present simplified versions of these results which will be used in this paper.

Proposition 2. *There exists a constant c such that for all n and k*

$$\sum_{i \leqslant (k-c)/a} \binom{n}{i} \leqslant \#A_{n,k} \leqslant \sum_{i \leqslant k/b} \binom{n}{i}. \tag{9}$$

[1] We suppose that some universal programming language is fixed. Let p be any program which given x and a degree of accuracy $\delta > 0$ computes a rational approximation of $S(x)$ with this accuracy. We denote by $K(S)$ the length of the shortest such p.

Proof. Let a sequence x of length n have no more than m ones. Consider prediction strategy $S(z) = 0$ for all z. Then by (6) and (7) there are at least $\sum_{i \leqslant m} \binom{n}{i}$ of x such that $\mathrm{KG}(x) \leqslant \mathrm{Loss}_S(x) + c \leqslant am + c \leqslant k$, where c is a constant. Then $m \leqslant (k - c)/a$ and we obtain the left-hand side of inequality (9).

To prove the right-hand side of inequality (9) consider the (possibly incomputable) *universal* prediction strategy $\Lambda(x) = p$, where $p = p(x)$ is the prediction from the item (i) of definition of the measure of predictive complexity. By definition $\mathrm{Loss}_\Lambda(x) \leqslant \mathrm{KG}(x)$ for each x. We consider a binary tree defined by the set Ξ of vertices with edges $(x, x0)$ and $(x, x1)$ for all $x \in \Xi$. By (5) for any x we have $\lambda(0, \Lambda(x)) \geqslant b$ or $\lambda(1, \Lambda(x)) \geqslant b$. By this property we assign new labelling to edges of the binary tree using letters A and B. We assign A to $(x, x0)$ and B to $(x, x1)$ if $\lambda(0, \Lambda(x)) \geqslant b$, and assign B to $(x, x0)$ and A to $(x, x1)$ otherwise. Evidently, two different sequences of length n have different labellings. For each edge $(x, x\sigma)$ labelled by A it holds $\lambda(\sigma, \Lambda(x)) \geqslant b$ and, hence, for any sequence x of length n having more than m As it holds $\mathrm{KG}(x) \geqslant \mathrm{Loss}_\Lambda(x) \geqslant bm$. Therefore, $\#A_{n,k} \leqslant \sum_{i \leqslant k/b} \binom{n}{i}$. □

Remember, that $K(x)$ is the Kolmogorov prefix complexity of x.

Proposition 3. *A positive constant c exists such that for all n and k such that $6b \leqslant k \leqslant \frac{bn}{2}$ inequality*

$$K(x) \leqslant \frac{3k}{b} \log \frac{n}{k/b} + c \tag{10}$$

holds for all x of the length n satisfying $\mathrm{KG}(x) \leqslant k$.

Proof. Let $k \leqslant bn/2$ and let $A_{n,k}$ be defined by (8). Since $\mathrm{KG}(x)$ is semicomputable from above, we have an algorithm enumerating all elements of $A_{n,k}$ given n and k. So, we can specify any $x \in A_{n,k}$ by n, k and the ordinal number of x in this enumeration. We use also the upper bound (9). This gives us a prefix method of encoding elements of $A_{n,k}$ given n and k. Therefore, we obtain

$$\begin{aligned} K(x) &\leqslant \log \sum_{i \leqslant k/b} \binom{n}{i} + \log n + \log k + c_1 \\ &\leqslant \log \frac{k}{b} \left(\frac{en}{k/b}\right)^{k/b} + 2 \log n + c_1 \leqslant \frac{3k}{b} \log \frac{n}{k/b} + c, \end{aligned}$$

where c_1 and c are positive constants and $k \geqslant 6b$. □

Notice, that requirement (7) to loss function did not used in the proof of this proposition.

3. Expanding property

The following Proposition 4 is an analog of the corresponding result for the prefix Kolmogorov complexity $K(x) \leqslant K(x|y) + K(y) + O(1)$ [5].

Proposition 4. *Positive constants c_1 and c_2 exist such that for all x and y of length n*

$$\mathrm{KG}(x) \leqslant \mathrm{KG}(x|y) + (\ln 2/\eta)K(y) + c_1 \tag{11}$$

$$\leqslant \mathrm{KG}(x|y) + c_2\,\mathrm{KG}(y)\log\left(\frac{n}{\mathrm{KG}(y)}\right). \tag{12}$$

Inequality (12) *holds if* $6b \leqslant \mathrm{KG}(y) \leqslant bn/2$.

Proof. This proposition is a direct corollary of Proposition 1. For any finite sequence y define $\bar{y} = y0^\infty$. It is easy to verify that a function $KS(x) = \mathrm{KG}(x|\bar{y}^{l(x)})$ is a measure of predictive complexity with respect to the same loss function (i.e. it satisfies requirements (i) and (ii) of the definition of the measure of predictive complexity). It can be enumerated from below by a program depending from y. Then by (4) for any x such that $l(x) = l(y)$ it holds $\mathrm{KG}(x) \leqslant \mathrm{KG}(x|y) + (\ln 2/\eta)K(y) + c_1$, where c_1 is a positive constant. Applying the upper bound on $K(y)$ from Proposition 3 we obtain the needed result. □

The following theorem shows that inequality (12) cannot be improved. We construct a computable mapping having some "expanding property": it transforms sequences of given predictive complexity into sequences of essentially bigger predictive complexity.

Let $C_{n,k}$ be the set of sequences y of the length n having k changes from 0 to 1 or from 1 to 0 (occurrences of combinations 01 and 10 in y); it is also convenient to consider a case $y_1 = 1$ (i.e. the case when y starts from 1) as a change.

Let us consider a computable predictive strategy S such that $S(\varepsilon) = 0$ and $S(x1) = 1$, $S(x0) = 0$ for all x. By (7) we have $\mathrm{Loss}_S(y) \leqslant a(k+1)$ for each $y \in C_{n,k}$. Therefore, $\mathrm{KG}(y) \leqslant ak + O(1)$ for each $y \in C_{n,k}$. Let $\lceil r \rceil$ be the least integer number s such that $s \geqslant r$.

For any n and any sequence z of the length $\leqslant n$ by subtree $\Xi_n(z)$ we mean the set

$$\Xi_n(z) = \{y | l(y) = n, z \subseteq y\}$$

supplied by corresponding binary tree structure. The height of the subtree is the number $m = n - l(z)$.

Theorem 1. *For any n and $k \leqslant \frac{n}{2}$ a computable mapping Φ from $C_{n,k}$ to the subtree $\Xi_n(0^{n-m})$, where $m = \lceil \log 4^k \binom{n}{k} \rceil$, exists such that for a portion $\geqslant \frac{1}{2}$ of all $y \in C_{n,k}$ the output $x = \Phi(y)$ satisfies*

$$\mathrm{KG}(x) \geqslant \frac{b}{20}\log\binom{n}{k}, \tag{13}$$

$$\mathrm{KG}(x|y) = O(\log n). \tag{14}$$

We have also $\mathrm{KG}(y) \leqslant ak + O(1)$ *for each* $y \in C_{n,k}$. *In the case* $k = k(n) = o(n)$ *the factor* $b/20$ *in* (13)–(14) *can be replaced on* $b/2$.

Proof. We construct a mapping Φ which compresses the set $C_{n,k}$ into the subtree $\Xi_n(0^{n-m})$ of height m, where $m = \lceil \log 4^k \binom{n}{k} \rceil$, such that the density of the image of Φ in this set is sufficiently large, namely, 4^{-k}. We prove also a variant of "incompressibility lemma", from which follows that the large portion of this image consists of elements x of predictive complexity $\mathrm{KG}(x)$ satisfying (13).

Let $y * y'$ denotes the maximal joint prefix of y and y', i.e. a sequence z of maximal length such that $z \subseteq y$ and $z \subseteq y'$. The mapping Φ will satisfy

$$l(\Phi(y) * \Phi(y')) \geqslant l(y * y') \tag{15}$$

for all $y, y' \in C_{n,k}$.

We construct the mapping Φ as a result of the recursive procedure COMP(n, k, σ), where n and k be positive integer numbers and $\sigma = 0$ or 1.

Procedure COMP(n, k, σ).

For any n the procedure COMP$(n, 0, 0)$ returns the identical mapping Φ on the set $\{0^n\}$, the procedures COMP$(n, 1, 1)$ and COMP$(n, 0, 1)$ return the identical mapping Φ on the set $\{1^n\}$.

Let $C^{\sigma}_{n,k}$ be the set of all sequences from $C_{n,k}$ starting from σ ($\sigma = 0, 1$).

Let $k > 0$. Then the procedure returns a mapping Φ from $C^{\sigma}_{n,k}$ into a subtree $\Xi_n(\sigma^l)$ for some $l \geqslant 1$. Without loss of generality we suppose that $\sigma = 0$. We will construct this mapping Φ by series of subsequent reassignments of the values of Φ. We start with the mapping Φ identical on $C^0_{n,k}$.

We distinguish two main stages of the construction.

Stage 1. *The preliminary definition of* Φ *on basic subtrees* (*recursive application of the procedure* COMP *to subtrees of the binary tree*).

There are $n-1$ *basic subtrees* $\Xi_n(0^i)$, $i = 1, \ldots, n-1$. Each of them is isomorphic (by a natural way) to the tree Ξ_{n-i}. For each $i = 1, \ldots, n-2$ we apply the procedure COMP$(n-i, k-1, 1)$ to the tree Ξ_{n-i}. By its definition the procedure returns for each i a mapping Φ_i from $C^1_{n-i,k-1}$ into a subtree $\Xi_{n-i}(1^{k_i})$ for some $k_i > 0$ (also $k_i < n-i$).

Induction hypothesis: for each i the set $\Phi_i(C^1_{n-i,k-1}) \cap \Xi_{n-i}(1^{k_i})$ occupies at least $4^{-(k-1)}$ portion of all sequences of subtree $\Xi_{n-i}(1^{k_i})$. In other words,

$$\#\Phi_i(C^1_{n-i,k-1}) \cap \Xi_{n-i}(1^{k_i}) \geqslant 4^{-(k-1)} \#\Xi_{n-i}(1^{k_i}).$$

We say also that the density of the image of Φ_i in $\Xi_{n-i}(1^{k_i})$ is at least $4^{-(k-1)}$. Any $x \in C^0_{n,k}$ can be represented as $x = 0^i 1z$ for some $i > 0$ and for some z of length $n-i-1$. Define $\Phi(x) = 0^i \Phi_i(1z)$.

Denote $A_i = \Xi_n(0^i 1^{k_i})$ and call it *compressed basic subtree*. The height of A_i is equal to $h_i = n-i-k_i$. A sequence of trees A_i is called *normal* if the inequality $h_i \geqslant h_{i+1}$ holds for all i.

Lemma 1. *A normal sequence of compressed basic subtrees* A_i *can be constructed.*

Proof. To construct this sequence we change the assignments of Φ according to the following instructions. Each initial basic subtree $\Xi_n(0^i)$, $1 \leqslant i < n-1$ contains (before applying the procedure COMP) a subtree $T = \Xi_n(0^i 1)$ isomorphic to its upper neighbour $\Xi_n(0^{i+1})$. This subtree T will be transformed by the procedure COMP as a part of $\Xi_n(0^i)$ on the previous inductive step into a subtree T' of corresponding resulting basic subtree $A_i = \Xi_n(0^i 1^{k_i})$. For each i if $h_i < h_{i+1}$ then replace A_{i+1} on the corresponding subtree T' of A_i. Doing these replacements we simultaneously changing corresponding assignments of the values of Φ. This process provides us that $h_i \geqslant h_{i+1}$ for all i. □

Let A_i be a normal sequence of compressed basic subtrees. By definition $\Phi(0^i 1z) \in A_i$ for all z of length $n-i-1$. By the induction hypothesis $\Phi(C^0_{n,k}) \cap (\cup_i A_i)$ occupies at least $4^{-(k-1)}$ portion of all sequences from $\bigcup_i A_i$.

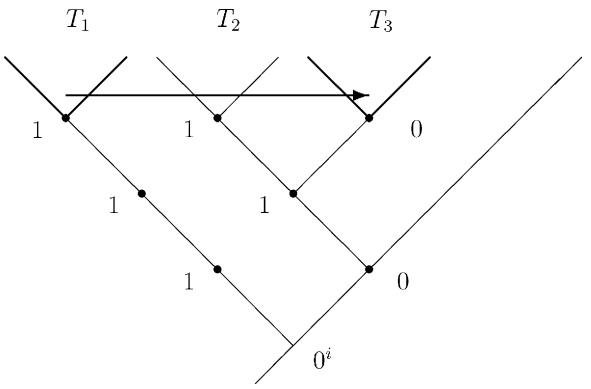

Fig. 1. Joining to the nearest upper neighbour ($s_i = 0$). The values of Φ located in the subtree $T_1 = \Xi_n(0^i 1^{s_i+3})$ are transformed to values located in the subtree $T_3 = \Xi_n(0^{i+1}1^{s_i+3}0)$.

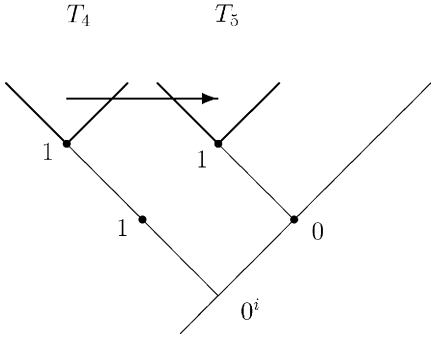

Fig. 2. Moving up to the nearest upper free vertex on the base ($s_i = 0$). The values of Φ located in the subtree $T_4 = \Xi_n(0^i 1^2)$ are transformed to values located in the subtree $T_5 = \Xi_n(0^{i+1}1)$.

Stage 2. *The final transformation of* Φ (*Joining and moving of the compressed basic subtrees to the resulting compressed subtree* $\Xi_n(0^l)$).

We will change the values of Φ defined at Stage 1 such that the density of the image of Φ in some resulting subtree $\Xi_n(0^l)$ will be at least 4^{-k}. The reconstruction consists of subsequent application of the following two operations on subtrees of a binary tree and of corresponding reassignments of Φ.

(1) *Joining to the nearest upper neighbour* (see Fig. 1). Suppose that there are at least two compressed basic subtrees (defined at Stage 1) of the type $T_1 = \Xi_n(0^i 1^{s_i+3})$ and $T_2 = \Xi_n(0^{i+1}1^{s_i+2})$, where $s_i \geqslant 0$, and of the same height h. We transform each assignment of the type $\Phi(y) = 0^i 1^{s_i+3}z$ to $\Phi(y) = 0^{i+1}1^{s_i+1}0z$, i.e. move subtree $T_1 = \Xi_n(0^i 1^{s_i+3})$ to a tree $T_3 = \Xi_n(0^{i+1}1^{s_i+1}0)$. We will apply this transformation only to the first (left) pair of the basic subtrees of height h. Then the height of the resulting subtree will be $h + 1$ and will not exceed the height of the previous basic subtree.

(2) *Moving up to the nearest upper free vertex on the base* (see Fig. 2). Suppose that after Stage 1 there is a basic subtree $T_4 = \Xi_n(0^i 1^{s_i+2})$ and there is no basic subtree of type $T_5 = \Xi_n(0^{i+1}1^{s_i+1})$, where $s_i \geqslant 0$. In this case we change each assignment $\Phi(y) = 0^i 1^{s_i+2}z$ to $\Phi(y) = 0^{i+1}1^{s_i+1}z$, we say that the basic subtree $T_4 = \Xi_n(0^i 1^{s_i+2})$ is moved to a subtree $T_5 = \Xi_n(0^{i+1}1^{s_i+1})$.

By definition, after these transformations, the image of Φ still occupies at least $4^{-(k-1)}$ portion of all sequences from each transformed basic subtree.

Given initial compressed basic subtrees we transform them by applying these two operations as long as they output nontrivial results (we omit the formal description of the corresponding algorithm).

Lemma 2. *Any basic subtree has the form $\Xi_n(0^l1)$ or $\Xi_n(0^l11)$ after these transformations, where $l \geqslant 1$.*

Proof. Suppose that among final basic subtrees there is a basic subtree of type $\Xi(0^i1^{s_i+3})$, where $s_i \geqslant 0$. Then there are only two possibilities. The first one is that there is no basic subtree of type $\Xi_n(0^{i+1}1^{s_i+2})$, and then the operation (2) must be applied. The second one is that there is a basic subtree $\Xi_n(0^{i+1}1^{s_i+2})$ (and so, subtree $\Xi_n(0^{i+1}1^{s_i+1}0)$ has empty intersection with the image of Φ, since otherwise subtree $\Xi_n(0^{i+1}1^{s_i})$ should be basic). In this case the operation (1) must be applied. No basic subtree of type $\Xi_n(0^{i+1}1^s)$ (where $s \leqslant s_i+1$) was constructed on previous step of recursion, since otherwise, the property $h_i \geqslant h_{i+1}$ of the normality of compressed basic subtrees (see Lemma 1) fails. The contradiction obtained shows that there is no basic subtree of type $\Xi(0^i1^{j+3})$ after all transformations. □

As the result of Stage 2 we obtain new assignments to values of Φ. These new values are contained in new basic subtrees of the type $\Xi_n(0^i1^j)$, where $i \geqslant 1$ and $j = 1$ or 2. We note also that new assignments never destroy the property (15) of Φ.

Consider the first basic subtree B of maximal height. Since, the operations (1) and (2) cannot be applied to B, then by Lemma 2, it is of the form $\Xi_n(0^l1)$ or $\Xi_n(0^l11)$, where $l \geqslant 1$. Then the portion of sequences of B among all sequences of $\Xi_n(0^l)$ is at least $\frac{1}{4}$. Therefore, the portion of sequences in $\Xi_n(0^l)$, which are of a form $\Phi(y)$, where $y \in C^0_{n,k}$, can decrease no more than in 4 times, and so it is $\geqslant \frac{1}{4}4^{-(k-1)} = 4^{-k}$.

We declare the mapping Φ and the subtree $\Xi_n(0^l)$ as the results returning by procedure COMP$(n, k, 0)$. We also proved that the induction hypothesis holds for these results.

End of the procedure COMP.

Let Φ and $\Xi_n(0^l)$ be outputs of COMP$(n, k, 0)$. Since the set $\Phi(C_{n,k})$ contains $\binom{n}{k}$ sequences in the binary tree $\Xi_n(0^l)$ by density condition this tree has $\leqslant 4^k\binom{n}{k}$ sequences.

The following lemma will be used to prove the existence of elements of big predictive complexity in a set of given cardinality.

Lemma 3. *For any n let $m < n$ and z be a sequence of the length $n - m$. Then for any set $W \subseteq \Xi_n(z)$ for at least $\frac{1}{2}$ portion of all $x \in W$ it holds*

$$\mathrm{KG}(x) \geqslant bl_{\max},$$

where $l_{\max}$ is the maximal integer number l such that $l \leqslant \frac{m}{2}$ and

$$H\left(\frac{l}{m}\right) \leqslant \frac{\log \#W}{m} - \frac{\log m}{m} - \frac{1}{m}, \tag{16}$$

where $H(p) = -p \log p - (1-p)\log(1-p)$ is the Shannon entropy.

Proof. Let $\Lambda(x)$ be the universal predictive strategy as in the proof of Proposition 2 such that $\mathrm{Loss}_\Lambda(x) \leqslant \mathrm{KG}(x)$ for all x. By definition of b (see (5)) for any x we have $\lambda(0, \Lambda(x)) \geqslant b$ or $\lambda(1, \Lambda(x)) \geqslant b$. Using this

property we assign the labelling to edges using letters A and B as in the proof of Proposition 2. Then for any sequence x of the length m having more than k As it holds $\mathrm{Loss}_{\Lambda}(x) \geqslant bk$.

Now, to estimate from below the maximal total loss of Λ on sequences from W we must estimate from below the maximal number of As occurring in sequences from W. This estimate l must satisfy inequality

$$\sum_{i=1}^{l} \binom{m}{i} < \frac{1}{2} \#W. \tag{17}$$

We have $\mathrm{Loss}_{\Lambda}(x) \geqslant bl$ for all x from the at least $\frac{1}{2}$ part of elements x of W (where the number of As in x is more than l). Inequality (17) follows from $l\binom{m}{l} < \frac{1}{2}\#W$. Since the elementary inequality

$$\binom{m}{l} \leqslant 2^{mH\left(\frac{l}{m}\right)}$$

holds (see [2, Section 6.1]), inequality (17) also follows from:

$$l2^{mH\left(\frac{l}{m}\right)} \leqslant \frac{1}{2} \#W. \tag{18}$$

This inequality follows from:

$$H\left(\frac{l}{m}\right) \leqslant \frac{\log \#W}{m} - \frac{\log m}{m} - \frac{1}{m}. \tag{19}$$

Hence, we have $\mathrm{KG}(x) \geqslant bl$ for at least $\frac{1}{2}$ portion of all $x \in W$, where l is the maximal number satisfying (19). □

To apply Lemma 3 get

$$m = \left\lceil \log 4^k \binom{n}{k} \right\rceil,$$

$$W = \Phi(C_{n,k}).$$

We have $\#W = \binom{n}{k}$. We must find maximal l such that the inequality

$$H\left(\frac{l}{m}\right) \leqslant \frac{\log \binom{n}{k}}{2k + \log \binom{n}{k}} - \frac{\log m}{m} - \frac{1}{m} \tag{20}$$

holds. It holds $\binom{n}{k} \geqslant 2^k$ if $k \leqslant \frac{n}{2}$. Then since $\log \binom{n}{k} \geqslant k$, the first term of (20) is bigger than $\frac{1}{3}$. Hence, for m sufficiently large it is sufficient to find l such that

$$H\left(\frac{l}{m}\right) \leqslant \frac{3}{10}. \tag{21}$$

It is easy to verify by table values of Shannon entropy that inequality (21) wittingly holds if

$$\frac{l}{m} \leqslant \frac{1}{20}$$

and so, we can take estimate

$$l_{\max} = \frac{m}{20}.$$

By Lemma 3 an x of the length n exists such that

$$\mathrm{KG}(x) \geqslant \frac{b}{20}\left(2k + \log\binom{n}{k}\right) \geqslant \frac{b}{20}\log\binom{n}{k}.$$

Let $\Phi(y) = x$. By the prefix property (15) of mapping Φ each prefix x^i of length i can be recovered from a prefix y^i of length i. Hence, a prediction strategy S exists which computes the ith member of x given $y^i = y_1 \ldots y_i$. By definition we have $\mathrm{Loss}_S(x|y) = 0$. This predictive strategy S is trivially defined by the mapping Φ given n and k as parameters. Hence, by (4) we have $\mathrm{KG}(x|y) \leqslant (2\ln 2/\eta)\log n$ for all sufficiently large n. The proof of the theorem is completed. □

In the following theorems we present results of Proposition 4 and Theorem 1 in an asymptotic form. For any functions $\alpha(n)$ and $\beta(n)$ the expression $\alpha(n) = \Theta(\beta(n))$ means that there exist constants c_1 and c_2 such that $c_1\beta(n) \leqslant \alpha(n) \leqslant c_2\beta(n)$ holds for all n. The expression $\alpha(n) = \Omega(\beta(n))$ means that there exists a constant c_1 such that $\alpha(n) \geqslant c_1\beta(n)$ for all n.

Theorem 2. *Let a function $k(n)$ be unbounded, $v(n) = \Omega(\log n)$, and $k(n) = O(n)$, $v(n) = O(n)$. Then*

$$\sup_{x:\exists y(\mathrm{KG}(y)\leqslant k(n),\mathrm{KG}(x|y)\leqslant v(n))} \mathrm{KG}(x) = \Theta\left(v(n) + k(n)\log\left(\frac{n}{k(n)}\right)\right). \tag{22}$$

Proof. The part $\leqslant$ follows from Propositions 3 and 4. In case $v(n) \geqslant k(n)\log\left(\frac{n}{k(n)}\right)$ the part $\geqslant$ of (22) is evident. Otherwise, the statement follows from Theorem 1. □

For any $x = x_1 \ldots x_n$ and $y = y_1 \ldots y_n$ we consider "a pair" $[y, x] = y_1x_1 \ldots y_nx_n$.

Theorem 3. *Let $k(n) = \Omega(\log n)$, $v(n) = \Omega(k(n))$, $k(n) = O(n)$ and $v(n) = O(n)$. Then*

$$\sup_{x:\exists y(\mathrm{KG}(y)\leqslant k(n),\mathrm{KG}([y,x])\leqslant v(n))} \mathrm{KG}(x) = \Theta\left(v(n) + k(n)\log\left(\frac{n}{k(n)}\right)\right). \tag{23}$$

Proof. In case $v(n) \geqslant k(n)\log\left(\frac{n}{k(n)}\right)$ the part $\geqslant$ of (23) is evident. Otherwise, we use Theorem 1 and its proof. Let us consider the mapping Φ from the proof of Theorem 1 and the sequences y and $x = \Phi(y)$ satisfying the conditions of this proposition. Define a computable prediction strategy S by $S(y_1x_1 \ldots y_{i-1}x_{i-1}y_i) = \Phi(y^i)_i = x_i$, (see also the end of the proof of Theorem 1), and $S(y_1x_1 \ldots y_ix_i) = 0$. This strategy predicts each even bit precisely and outputs 0 as prediction for any odd bit. Then by (7) we have $\mathrm{Loss}_S([y, x]) \leqslant ak(n)$ and, therefore,

$$\mathrm{KG}([y, x]) \leqslant ak(n) + (\ln 2/\eta)K(S) \leqslant ak(n) + (2\ln 2/\eta)\log n \leqslant c_1v(n) \tag{24}$$

for some constant c_1. By Theorem 1 it holds $\mathrm{KG}(y) \leqslant c_2 \log n \leqslant c_3k(n)$ for all y, where c_2, c_3 are constants. The part $\geqslant$ of (23) follows from these inequalities after normalizing of $k(n)$ and $v(n)$.

The $\leqslant$ part of (23) follows from Proposition 4 and an obvious inequality $\mathrm{KG}(x|y) \leqslant \mathrm{KG}([y, x]) + c$, where c is a positive constant. $\square$

4. Predictive information

The *amount of predictive information* in a sequence y about a sequence x of the same length in the process of on-line prediction was defined by Vovk [9]

$$\mathrm{IG}(y : x) = \mathrm{KG}(x) - \mathrm{KG}(x|y). \tag{25}$$

In this section, we explore relations between four important values $\mathrm{IG}(y : x)$, $\mathrm{IG}(x : y)$, $\mathrm{KG}(x)$ and $\mathrm{KG}(y)$ in a limit form. These results are mainly based on the construction of Theorem 1.

Theorem 4 below implies that predictive information is noncommutative in the strongest possible sense. Define

$$g_1(n) = \sup_{l(x)=l(y)=n} (\mathrm{IG}(x : y) - \mathrm{IG}(y : x)). \tag{26}$$

Theorem 4. *It holds* $g_1(n) = bn + O(1)$.

Proof. For any sequence $x = x_1 \dots x_n$ define its left shift $Tx = x_2 \dots x_n 0$. The proof of theorem is based on the following simple lemma.

Lemma 4. *It holds*

$$\mathrm{KG}(x|Tx) = O(1), \tag{27}$$

$$\mathrm{KG}(Tx) = \mathrm{KG}(Tx|x) + O(1). \tag{28}$$

To prove (27) consider a computable prediction strategy S such that for any sequences x and y of length $n \geqslant 2$ it holds $S(x|y) = y_{n-1}$ (for $n = 1$ define $S(x|y) = 0$). Then $\mathrm{Loss}_S(x|Tx) \leqslant b$ for each x, and by definition $\mathrm{KG}(x|Tx) \leqslant b + c$ hold for all x, where c is a positive constant c.

Let us prove (28). We have $\mathrm{KG}(Tx|x) \leqslant \mathrm{KG}(Tx) + O(1)$ by definition. To prove the converse inequality define a function $S(u)$ using an idea of Vovk's [8] aggregating algorithm.

$$S(u) = \log_\beta(2^{-1}\beta^{\mathrm{KG}(u|0u)} + 2^{-1}\beta^{\mathrm{KG}(u|1u)}). \tag{29}$$

We show that $S(u)$ is a measure of predictive complexity. Indeed, for any σ

$$\begin{aligned} S(u\sigma) - S(u) &= \log_\beta \sum_{i=0}^{1} 2^{-1}\beta^{\mathrm{KG}(u\sigma|iu)} - \log_\beta \sum_{i=0}^{1} 2^{-1}\beta^{\mathrm{KG}(u|iu)} \\ &= \log_\beta \sum_{i=0}^{1} q_i \beta^{\mathrm{KG}(u\sigma|iu) - \mathrm{KG}(u|iu)} \geqslant \log_\beta \sum_{i=0}^{1} 2^{-1}\beta^{\lambda(\sigma, \hat{p}(iu))} \geqslant \lambda(\sigma, \hat{p}(u)), \end{aligned}$$

where

$$q_i = \frac{2^{-1}\beta^{\mathrm{KG}(u|iu)}}{\sum_{j=0}^{1} 2^{-1}\beta^{\mathrm{KG}(u|ju)}},$$

$i = 0, 1$, and predictions $\hat{p}(iu)$ and $\hat{p}(u)$ exist by η-mixability property of the loss function $\lambda(\sigma, p)$. By definition (29) of $S(u)$ we have for $i = 0, 1$

$$S(u) \leqslant \mathrm{KG}(u|iu) + (\ln 2/\eta)$$

for all u. Hence, by definition

$$\mathrm{KG}(Tx) \leqslant \mathrm{KG}(Tx|x) + c$$

for all x, where c is a positive constant. □

To prove the inequality $g_1(n) \geqslant bn + O(1)$ of Theorem 4 let us compare $\mathrm{IG}(Tx : x)$ and $\mathrm{IG}(x : Tx)$ for some x. By Lemma 4 a positive constant c exists such that

$$\mathrm{IG}(Tx : x) = \mathrm{KG}(x) - \mathrm{KG}(x|Tx) \geqslant \mathrm{KG}(x) - c,$$

$$\mathrm{IG}(x : Tx) = \mathrm{KG}(Tx) - \mathrm{KG}(Tx|x) \leqslant c$$

for all x. Using definition (5) of b and the diagonal argument it easy to construct a sequence x of length n such that $\mathrm{KG}(x) \geqslant bn$. From this the inequality $g_1(n) \geqslant bn + O(1)$ follows.

The inequality $g_1(n) \leqslant bn + O(1)$ follows from inequalities:

$$\mathrm{IG}(x : y) - \mathrm{IG}(y : x) \leqslant \mathrm{KG}(y) \leqslant bl(y) + c$$

for all x and y, where c is a positive constant. The last inequality can be easily proved using a computable prediction strategy which always predicts $\hat{p}$, where $\hat{p}$ minimizes condition (5). □

Let us define

$$g_2(n) = \sup_{l(x)=l(y)=n, 6b \leqslant \mathrm{KG}(y)} \frac{\mathrm{IG}(y : x)}{\mathrm{KG}(y)}. \tag{30}$$

The main results of Section 3 can be summarized in the following theorem.

Theorem 5. *It holds* $g_2(n) = \Theta(\log n)$.

Proof. Inequality $g_2(n) \leqslant C_1 \log n$ (for some $C_1 > 0$) follows directly from (12). Inequality $g_2(n) \geqslant C_2 \log n$ (for some $C_2 > 0$) can be derived from Theorem 1. Its enough to let $k = \sqrt{n}$. □

Let us define also

$$g_3(n) = \sup_{l(x)=l(y)=n, 0 < \mathrm{KG}(x)} \frac{\mathrm{IG}(y : x)}{\mathrm{KG}(x)}. \tag{31}$$

Theorem 6. *It holds* $\lim_{n\to\infty} g_3(n) = 1$.

Proof. This relation follows from the representation:

$$g_3(n) = \sup_{l(x)=l(y)=n, 0<\mathrm{KG}(x)} \left(1 - \frac{\mathrm{KG}(x|y)}{\mathrm{KG}(x)}\right)$$

and from Theorem 1, where we let $k = n/4$. □

Theorem 7. *Let $1 \leqslant k(n) \leqslant bn$ for all n. Then a constant c exists such that*

$$\sup_{(x,y):l(x)=l(y)=n, \mathrm{KG}(y)\leqslant k(n)} \mathrm{IG}(y:x) = \Theta\left(k(n)\log\left(\frac{n}{k(n)}\right)\right) \tag{32}$$

$$\sup_{(x,y):l(x)=l(y)=n, \mathrm{KG}(x)\leqslant k(n)} \mathrm{IG}(y:x) = k(n) + O(1) \tag{33}$$

$$\sup_{(x,y):l(x)=l(y)=n, \mathrm{IG}(x:y)\leqslant c} \mathrm{IG}(y:x) = \Theta(n). \tag{34}$$

Proof. The part $\leqslant$ of (32) follows from Proposition 4. The part $\geqslant$ of (32) follows from Theorem 1. To prove (33) put $x = y$ and note that $\mathrm{KG}(x|x) = O(1)$. Relation (34) follows from the proof of Theorem 4. □

Appendix A. Proof of Proposition 1

A sequence $\mathrm{KG}_i(x|y)$ of all measures of predictive complexity (with respect to a mixable loss function $\lambda(\sigma, \gamma)$) can be defined using standard methods of the theory of algorithms (see [12]). Let r_i be a semi-computable from below sequence of real numbers such that the series $\sum_{i=1}^{\infty} r_i$ is convergent and its sum does not exceed 1. We can take $r_i = 2^{-K(i)}$. Analogously to [9,10] a measure of predictive complexity $\mathrm{KG}(x|y)$ can be defined

$$\mathrm{KG}(x|y) = \log_\beta \sum_{i=1}^{\infty} \beta^{\mathrm{KG}_i(x|y)} r_i, \tag{A.1}$$

where $\beta = e^{-\eta}$. By definition $\mathrm{KG}(x|y)$ is semicomputable from above, i.e (ii) holds. We must verify (i). Indeed, by (A.1) for every x, y of equal length and $\sigma, \beta \in \{0, 1\}$

$$\mathrm{KG}(x\sigma|y\beta) - \mathrm{KG}(x|y) = \log_\beta \sum_{i=1}^{\infty} q_i \beta^{\mathrm{KG}_i(x\sigma|y\beta) - \mathrm{KG}_i(x|y)} \tag{A.2}$$

$$\geqslant \log_\beta \sum_{i=1}^{\infty} q_i \beta^{\lambda(\sigma,\gamma_i)} \geqslant \lambda(\sigma, \gamma), \tag{A.3}$$

where

$$q_i = \frac{r_i \beta^{\mathrm{KG}_i(x|y)}}{\sum_{s=1}^{\infty} r_s \beta^{\mathrm{KG}_s(x|y)}}.$$

Here for any i a prediction $\gamma_i = \gamma(x, y\beta)$ satisfying

$$\mathrm{KG}_i(x\sigma|y\beta) - \mathrm{KG}_i(x|y) \geqslant \lambda(\sigma, \gamma_i)$$

exists since each element of the sequence $\mathrm{KG}_i(x|y)$ satisfies the condition (i) of the measure of predictive complexity. A prediction γ satisfying (A.3) exists by mixability of the game. For further details see [9, Section 7.6].

Inequality (4) is an easy consequence of (3). To prove the rest part of Proposition 1 note that there exists a computable function $f(p)$ which transform any program p computing S into an enumerating program $i = f(p)$ such that $\mathrm{Loss}_S(x|y) = \mathrm{KG}_i(x|y)$. We have also $K(i) = K(f(p)) \leqslant K(p) + c$, where c is a constant.

Acknowledgements

This work is based on conference paper [13]. Authors are grateful to Volodya Vovk for useful discussions and for his suggestions about formulating of main results of the paper. The first author was supported by the EPSRC Grant GR/R46670/01. The second author was supported by RFBR Grant 01-01-01028.

References

[1] N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, D. Haussler, R.E. Schapire, M.K. Warmuth, How to use expert advice, J. ACM 44 (1997) 427–485.
[2] H. Cormen, E. Leiserson, R. Rivest, Introduction to Algorithms, McGraw-Hill, New York, 1990.
[3] D. Haussler, J. Kivinen, M.K. Warmuth, Tight worst-case loss bounds for predicting with expert advice, Technical Report UCSC-CRL-94-36, University of California at Santa Cruz, revised December 1994, 1994. Short version in P. Vitányi (Ed.), Computational Learning Theory, Lecture Notes in Computer Science, vol. 904, Springer, Berlin, 1995, pp. 69–83.
[5] M. Li, P. Vitányi, An Introduction to Kolmogorov Complexity and its Applications, second ed., Springer, New York, 1997.
[6] H. Rogers, Theory of Recursive Functions and Effective Computability, McGraw-Hill, New York, 1967.
[7] V. Vovk, Aggregating strategies, in: M. Fulk, J. Case (Eds.), Proceedings of the 3rd Annual Workshop on Computational Learning Theory, Morgan Kaufmann, San Mateo, CA, 1990, pp. 371–383.
[8] V. Vovk, A game of prediction with expert advice, J. Comput. System Sci. 56 (1998) 153–173.
[9] V. Vovk, C.J.H.C. Watkins, Universal portfolio selection, Proceedings of the 11th Annual Conference on Computational Learning Theory, 1998, pp. 12–23.
[10] V. Vovk, A. Gammerman, Complexity estimation principle, Comput. J. 42 (N4) (1999) 318–322.
[11] V.V. V'yugin, Does snooping help? Theoret. Comput. Sci. 276 (2002) 407–415.
[12] M.V. Vyugin, V.V. V'yugin, On complexity of easy predictable sequences, Inform. Comput. 178 (2002) 241–252.
[13] M.V. Vyugin, V.V. V'yugin, Predictive complexity and information, Proceedings of the 15th International Conference on Computational Learning Theory—COLT'02, Lecture Notes on Artificial Intelligence, vol. 2375, Springer, Berlin, 2002, pp. 90–104.
[14] K. Yamanishi, Randomized approximate aggregating strategies and their applications to prediction and discrimination, in: Proceedings of the Eighth Annual ACM Conference on Computational Learning Theory, Assoc. Comput. Machinery, New York, 1995, pp. 83–90.
[15] A.K. Zvonkin, L.A. Levin, The complexity of finite objects and the algorithmic concepts of information and randomness, Russ. Math. Surv. 25 (1970) 83–124.